UNIVERSITY OF COPENHAGEN

Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances

Andersen, Mikkel Meyer; Mogensen, Helle Smidt; Eriksen, Poul Svante; Olofsson, Jill Katharina; Asplund, Maria; Morling, Niels

Published in:
Forensic Science International: Genetics

DOI:
10.1016/j.fsigen.2013.01.005

Publication date:
2013

Citation for published version (APA):
Andersen, M. M., Mogensen, H. S., Eriksen, P. S., Olofsson, J. K., Asplund, M., & Morling, N. (2013). Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances. *Forensic Science International: Genetics*, *7*(3), 327-336. https://doi.org/10.1016/j.fsigen.2013.01.005

Download date: 05. May. 2023

Forensic Science International: Genetics 7 (2013) 327–336

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances

Mikkel Meyer Andersen [a,*], Helle Smidt Mogensen [b,1], Poul Svante Eriksen [a,2], Jill Katharina Olofsson [b,1], Maria Asplund [b,3], Niels Morling [b,4]

[a] Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark
[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, DK-2100 Copenhagen East, Denmark

ARTICLE INFO

Article history:
Received 15 June 2012
Received in revised form 17 January 2013
Accepted 28 January 2013

Keywords:
Signal strength
Truncated normal distribution
Logistic regression
Numerical likelihood optimisation
Bayesian information criterion (BIC)
Bootstrap validation

ABSTRACT

Y chromosome short tandem repeats (Y-STRs) are valuable genetic markers in certain areas of forensic case-work. However, when the Y-STR DNA profile is weak, the observed Y-STR profile may not be complete – i.e. locus drop-out may have occurred. Another explanation could be that the stain DNA did not have a Y-STR allele that was detectable with the method used (the allele is a 'null allele'). If the Y-STR profile of a stain is strong, one would be reluctant to consider drop-out as a reasonable explanation of lack of a Y-STR allele and would maybe consider 'null allele' as an explanation. On the other hand, if the signal strengths are weak, one would most likely accept drop-out as a possible explanation.

We created a logistic regression model to estimate the probability of allele drop-out with the Life Technologies/Applied Biosystems AmpF*l*STR® Yfiler® kit such that the trade-off between drop-outs and null alleles could be quantified using a statistical model. The model to estimate the probability of drop-out uses information about locus imbalances, signal strength, the number of PCR cycles, and the fragment size of Yfiler. We made two temporarily separated experiments and found no evidence of temporal variation in the probability of drop-out. Using our model, we found that for 30 PCR cycles with a 150 bp allele, the probability of drop-out was 1:5000 corresponding to the average estimate of the probability of Y-STR null alleles at a signal strength of 1249 RFU. This means that the probability of a null allele is higher than that of an allele drop-out at e.g. 4000 RFU and the probability of drop-out is higher than that of a null allele at e.g. 75 RFU.

© 2013 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Y chromosome short tandem repeats (Y-STRs) are valuable genetic markers in forensic case-work, especially in sexual assault cases where only small amounts of DNA from a male perpetrator is found in combination with a large amount of DNA from a female victim [1–3]. The reason for this is that the routine investigation of autosomal STRs, in such cases, will result in a DNA profile of the female victim, while investigations of Y-chromosome markers will result in a male Y-STR profile even if the amount of female DNA is more that 1000 times larger than that of male DNA [4]. The weight of the evidence of matching Y-STR DNA profiles from e.g. a scene of crime and a suspect may be estimated by likelihood principles [5,6]. The weight of the evidence is usually presented as a likelihood ratio (LR) of

$$\frac{Pr(\text{Y-STR profile}|\text{the DNA comes from the suspect})}{Pr(\text{Y-STR profile}|\text{the DNA comes from a random person not related to the suspect})}.$$

To be able to calculate this, one must have a sound estimate of the probability of observing the Y-STR profile among random individuals in the relevant population. This is a problem in itself [7–10,24]. The other part of the LR is the probability of the Y-STR profile under the assumption that it comes from the suspect. This is easy if the Y-STR profiles of the crime scene sample and the suspect are identical – the probability is 1. However, when the amount of Y-STR DNA is small and the Y-STR DNA profile is weak, the observed Y-STR profile may not be complete – i.e. locus drop-out may have occurred. This phenomenon is often considered of minor importance, and the lack of result from a locus is often ignored under the assumption that the phenomenon was due to locus drop-out. However, another explanation could be that the stain DNA did not have a Y-STR allele that was detectable with the method used – typically due to a SNP in the primer binding regions around the Y-STR [11,12]. The average frequency of such 'null alleles' is approximately 1:5000 = 0.02% (in release 39 of

* Corresponding author. Tel.: +45 99408860.
E-mail addresses: mikl@math.aau.dk (M.M. Andersen), helle.smidt@forensic.ku.dk (H.S. Mogensen), svante@math.aau.dk (P.S. Eriksen), jill.olofsson@forensic.ku.dk (J.K. Olofsson), maria.asplund@forensic.ku.dk (M. Asplund), niels.morling@forensic.ku.dk (N. Morling).
[1] Tel.: +45 35326212.
[2] Tel.: +45 99408868.
[3] Tel.: +45 35326778.
[4] Tel.: +45 35326115.

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.01.005

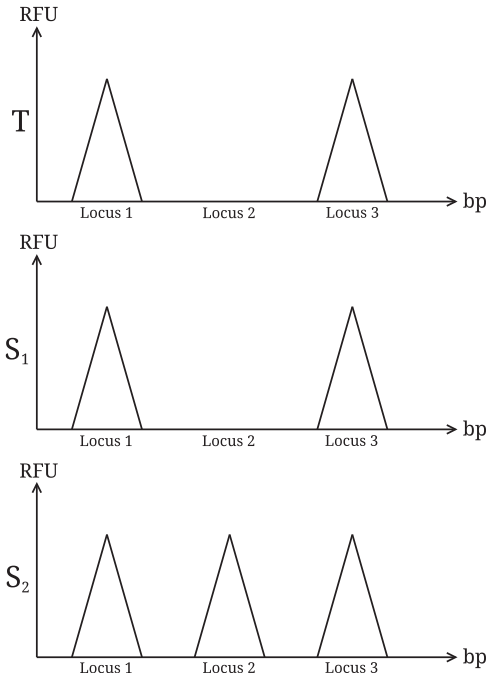

Fig. 1. An example that motivates to estimate the probability of allele drop-out. Assume that the topmost electropherogram (EPG) denoted by '*T*' was obtained from the evidence found at the crime scene and the two ones below are from two reference samples, 'S_1' and 'S_2'. Now, which reference sample is most consistent with '*T*'? 'S_1' can explain '*T*' by a null allele and 'S_2' can explain '*T*' by an allele drop-out. If the peaks in '*T*' are around e.g. 75 RFU, then we might suspect allele drop-out that would make 'S_2' consistent with '*T*'. On the other hand, if the peaks in '*T*' are around e.g. 4000 RFU, we would not suspect an allele drop-out, but instead suspect a null allele. Thus, in order to make a better analysis, we need a model to estimate the probability of allele drop-out compared to that of a null allele.

http://www.yhrd.org [13,14] there were 219 null alleles amount 1,111,984 alleles in total). If the Y-STR profile of a stain is strong with signal strength of e.g. 4000 RFU on an AB3130xl, drop-out is highly unlikely [15, own unpublished observations]. However, if the signal strength is e.g. 75 RFU, the probability of drop-out is approximately 20% (cf. Fig. 10), and drop-out must be included as a possible explanation.

Although the risk of drop-out may not seem so important for Y-STRs as for autosomal STRs [15,16], it should still be considered. We have investigated the drop-out risk of the AmpF*l*STR® Yfiler® (Life Technologies/Applied Biosystems) when using the kit with 28, 29, and 30 PCR cycles. We offer an easy method based on logistic regression analysis to estimate the drop-out risk of Y-STRs.

1.1. Motivating example

A simple example that motivates the evolution of the probability of allele drop-out is given in Fig. 1. A more complicated example is as follows: For the sake of argument, assume that the probability of a null allele at a locus is 1:5000 = 0.02% (which correspond to the number of null alleles in release 39 of http://www.yhrd.org [13,14]). Assume a two person mixture, where all but one locus has two peaks, each of height 4000 RFU. The last locus only has one peak of height 4000 RFU. The profile is well-balanced and there is no evidence of two shared alleles at this locus as this in theory would result in a peak of 8000 RFU. At 4000 RFU, the probability of drop-out is approximately 1:100,000 (cf. Table 1). This should be compared to the probability of a null allele (1:5000), which gives odds of 20 for a null allele compared to a drop-out.

Table 1
The signal strength to obtain a given drop-out probability at fragment sizes of 150 bp and 300 bp using a given number of PCR cycles. See Fig. 11 for a plot of this table.

P(Drop-out)	Bp	Cycles	Signal strength
0.001% (1:100,000)	150	28	1050
		29	1843
		30	4060
	300	28	1296
		29	2357
		30	5457
0.002% (1:50,000)	150	28	865
		29	1469
		30	3091
	300	28	1067
		29	1878
		30	4154
0.01% (1:10,000)	150	28	551
		29	867
		30	1640
	300	28	680
		29	1109
		30	2205
0.02% (1:5000)	150	28	453
		29	691
		30	1249
	300	28	560
		29	884
		30	1678
0.1% (1:1000)	150	28	289
		29	408
		30	663
	300	28	356
		29	522
		30	891
50% (1:2)	150	28	42
		29	43
		30	44
	300	28	51
		29	54
		30	59

Now, assume a two person mixture where all but one loci have two peaks, each of height 75 RFU. The last locus only has one peak of height 75 RFU. Again, we have a well-balanced profile where there is no evidence of two shared alleles at this locus. At 75 RFU, the probability of drop-out is approximately 1:5 (cf. Fig. 10). This should be compared to the probability of a null allele (1:5000), which gives odds of 1000 for a drop-out compared to a null allele.

2. Materials and methods

2.1. Experiments

Two sets of controlled experiments were conducted at The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark. For estimating the drop-out probability, eight different male DNA samples were diluted into 14 different concentrations and amplified in triplicates at 28, 29 and 30 thermocycles using the AmpF*l*STR® Yfiler® (Life Technologies/Applied Biosystems) amplification kit. The first set of experiments were conducted with DNA from four males. In the second set of experiments, DNA from four other males was investigated. In the first experiment, only data from 28 and 30 thermocycles were available.

For dilution series, blood samples were taken from eight males. Genomic DNA was extracted with the EZ1 Investigator kit (Qiagen) using a BioRobot EZ1 (Qiagen) or with PrepFiler™ Express Forensic

DNA Extraction Kit (AB) using an Automate Express™robot (AB). Each DNA sample was quantified in triplicate using the Quantifiler® Y Human Male DNA Quantification Kit (AB) with Human Genomic DNA Male (G147A, Promega) as the quantification standard on an ABIPrism 7000 (AB) or an ABIPrism 7500 (AB). The median DNA concentration was used. Each sample was diluted with water to DNA concentrations of 100 pg/μl or 1000 pg/μl. Dilution series were performed with serial dilutions to give 14 different DNA concentrations in the range 0.75–150 pg/μl.

A total of 5 or 10 μl of the diluted samples was added to the PCR mixture and each sample was amplified in triplicate with the AmpF*l*STR® Yfiler® PCR Amplification Kit (AB) as recommended by the manufacturer in an 96-Well GeneAmp® PCR System 9700 (AB) amplifying with 28, 29 and 30 thermocycles. The resulting amount of DNA in the PCR reactions ranged from 7.5 to 1000 pg.

One microliter of the amplificate together with 15 μl HiDi Formamide (AB) was analysed on an ABI Prism 3130xl Genetic Analyzer (AB) using POP4 (AB) as the polymer and 3 kV injection voltage for 10 s. DNA fragments were detected, fragment sizes were estimated, and alleles were assigned using GeneMapper 3.2 (AB) or GeneScan 3.7 with GenoTyper 3.7 (both AB) with a detection threshold of 15 RFU and no filter applied. A detection threshold of 50 RFU was used, which is also the detection threshold for drop-out. Peaks between 15 RFU and 50 RFU were included for improving statistical modelling.

The DNA profiles included only one allele per locus except for the DYS385a/b locus. Seven profiles had two alleles, and a single profile had one allele at the DYS385a/b locus.

The protocols were approved by the Danish ethical committee (KF-01-037/93 and H-1-2011-081).

2.2. Data

All data analysis was performed using the statistical software `R` [17].

In Fig. 2, the proportion of dropped out Y-STR loci given the expected DNA concentration and the number of PCR cycles for the sample is shown. In Fig. 3, the experiment is also included as a dependent variable.

No drop-out occured when the expected DNA concentration was greater than 100 pg/μl, which is why concentrations higher than 100 pg/μl are not shown in Figs. 2 and 3.

2.3. Estimating interlocus balances

The AmpF*l*STR Yfiler amplification kit is not well balanced between loci, which is depicted in Fig. 4. This means that locus balances need to be considered in the drop-out model. In this section, a model for estimating interlocus balances is described.

Due to the lack of accuracy and reproducibility in quantification, we could not use the quantified DNA amount in the model of the signal strength. Instead, we introduced an individual signal strength for each sample denoted by S_i for samples $i = 1, 2, \ldots, n$. The signal strength can be described as the mean peak height weighted by the interlocus balances. We will now discuss the modelling of this in detail.

Let x_{ij} be the peak height at the jth locus for the ith sample for $j = 1, 2, \ldots, r$ and $i = 1, 2, \ldots, n$, where r is the number of loci and n is the number of samples. Then, we assume that $\log x_{ij}$ is normally distributed with a mean value depending on the sample and locus. In a statistical notation, where $N(\mu, \sigma^2)$ denotes a normal distribution with a mean value μ and the variance σ^2, we assume that

$$\log x_{ij} \sim N(\theta_j + \log S_i, \sigma^2), \tag{1}$$

where θ_j is the locus balance for the jth locus and S_i is the signal strength for the ith sample.

We impose constraints on the θ_j's such that

$$\sum_{j=1}^{r} \theta_j = 0. \tag{2}$$

As the linear model stated in Eq. (1) assuming Eq. (2) is a linear regression model, we checked it on samples with full profiles (samples with no drop-out) using the linear model fit function `lm` in the statistical software `R` [17]. The adjusted R^2 value was 93.7% with both locus and sample as statistically significant factors. The resulting interlocus balances, θ_j, are depicted in Fig. 5.

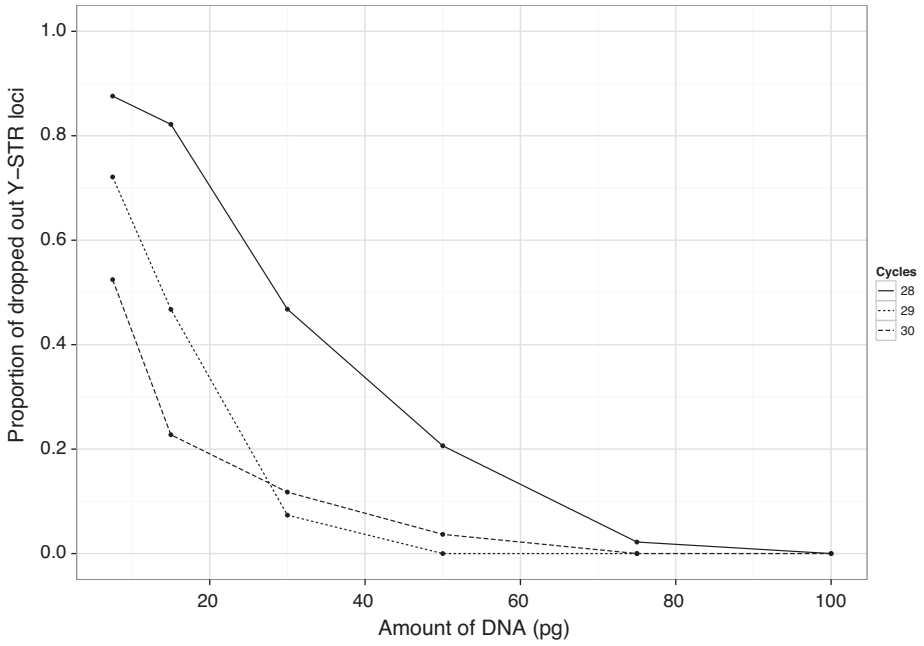

Fig. 2. The proportion of dropped out Y-STR loci depending on the amount of DNA. No drop-out occurred when the amount of DNA was greater than 100 pg.

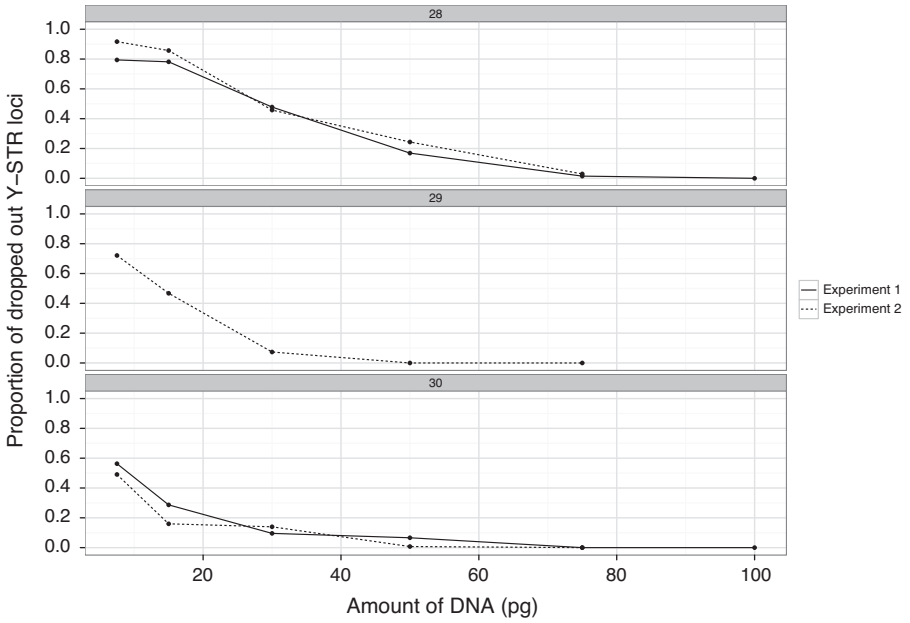

Fig. 3. The proportion of dropped out Y-STR loci given the amount of DNA, cycles and experiment. No drop-out occurred when the amount of DNA was greater than 100 pg.

For locus DYS385a/b, only one locus balance is estimated based on the sum of the peak heights of 2 alleles (7 profiles) and the peak height for 1 allele (1 profile). Later, for signal strength estimation, DYS385 was treated as two loci, 'DYS385a' and 'DYS385b', each with locus balance $\theta' = \theta/2$, where θ is this estimated locus balance for the sum of the DYS385a/b peak heights.

2.4. Estimating signal strength

Other studies on drop-outs, e.g. [15,16], use the signal strength as a predictor of the drop-out probability. We investigate the same predictor here. Due to the lack of balance of the Yfiler kit as described in Section 2.3, the signal strength must be modelled somewhat differently. Another difference in the modelling is that we incorporate the knowledge that some of the peaks may have dropped out by using a truncated probability distribution.

When we estimated interlocus balances on full profiles, we used the model in Eq. (1), Section 2.3. Now, when we have drop-outs, a slightly different model for the peak heights was used instead, namely

$$\log x_{ij} \sim N_{\log t}(\theta_j + \log S_i, \sigma_i^2), \tag{3}$$

where $N_{\log t}(\cdot\,,\,\cdot)$ denotes a normal distribution truncated below $\log t$ (meaning that there is no observation less than $\log t$, where t is

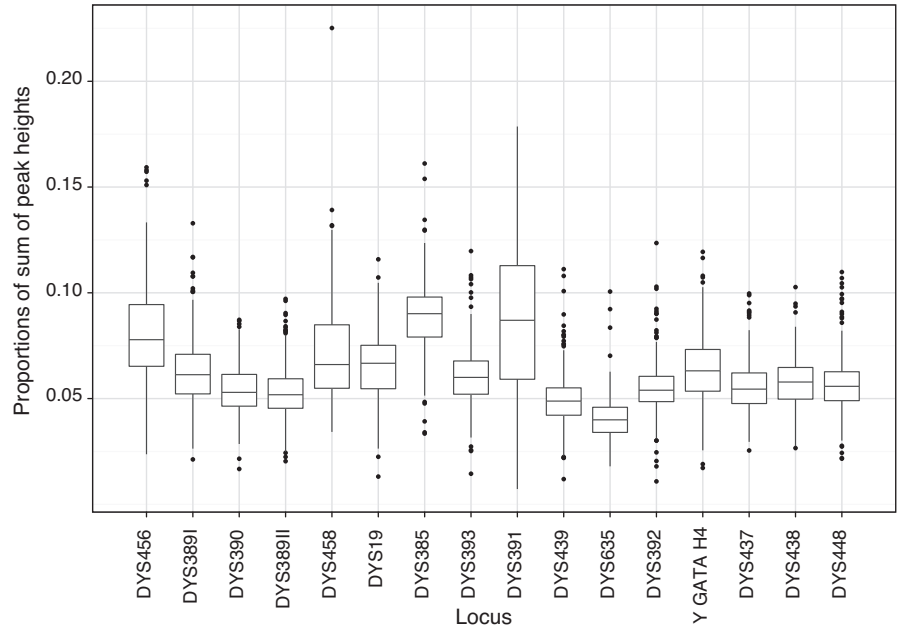

Fig. 4. Interlocus balances of the peak heights at the Y-STR loci. To explain the box-and-whiskers plot, let q_p be the p% quantile. The box contains the middle 50% of the observations (from the 25% quantile, q_{25}, to the 75% quantile, q_{75}). The horizontal line in the box displays the median (50% quantile, q_{50}). The end of the lower whisker is the lowest datapoint greater than $q_{25} - 1.5 \times$ IQR, where IQR is the interquartile range given by $q_{75} - q_{25}$ (the height of the box). The end of the upper whisker is the greatest datapoint lower than $q_{75} + 1.5 \times$ IQR. The points are outliers that are either lower than $q_{25} - 1.5 \times$ IQR or greater than $q_{75} + 1.5 \times$ IQR.

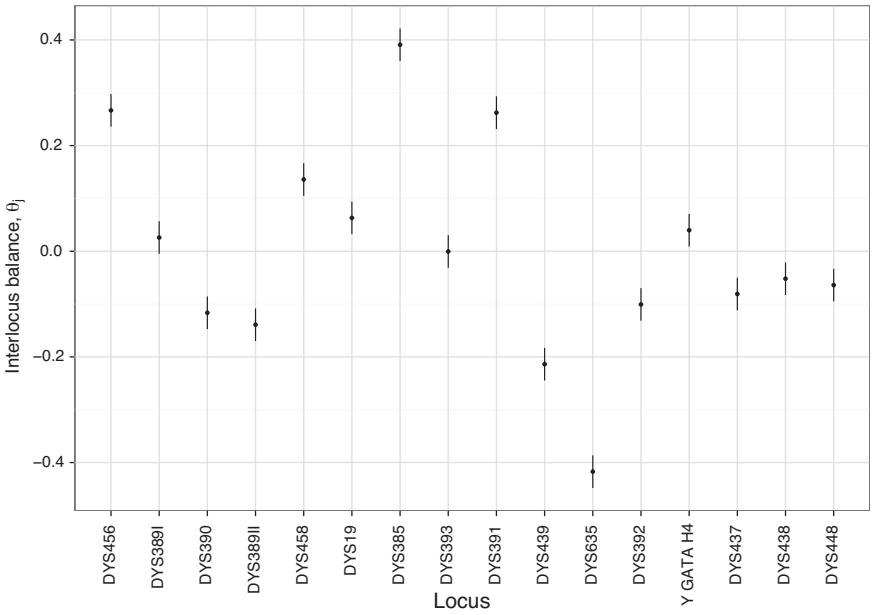

Fig. 5. Interlocus balances, θ_j, from the model $\log x_{ij} \sim N(\theta_j + \log S_i, \sigma^2)$ with 95% confidence intervals. Note, that all interlocus balance estimates have the same variance due to the balanced design (all samples are full profiles).

known and we have information about the number of observations being truncated). In forensic genetics, t is the detection threshold. Often the value t = 50 RFU is used, which we also used. As before, x_{ij} is the peak height at the jth locus and the ith sample, θ_j is the locus balance for the jth locus and S_i is the signal strength for the ith sample.

Now, assume that the interlocus balances estimated using Eq. (1) are known. This is a reasonable assumption and it makes inference about the signal strength, S_i, easier.

The goal is to estimate S_i and use it as a proxy for the signal strength by using the peaks above 50 RFU, their heights and implicitly peaks that have dropped out.

If we assume that the interlocus balances, θ_j, are known, then the model for one sample is

$$\log x_j \sim N_{\log t}(\theta_j + \log S, \sigma^2). \tag{4}$$

Let $J \subseteq \{1, 2, \ldots, r\}$ denote the set of loci that did not drop out and $J^C = \{1, 2, \ldots, r\} \setminus J$, where $\setminus$ means set difference, the set of loci that dropped out. The likelihood of the model in Eq. (3) for one sample $\{x_j\}_{j\in J}$ is then given by

$$\begin{aligned} L(\log S, \sigma^2; \{x_j\}_{j\in J}) &= \prod_{j=1}^{r} L_j \\ &= \prod_{j\in J^C} \Phi\left(\frac{\log t - (\theta_j + \log S)}{\sigma}\right) \\ &\quad \times \prod_{j\in J} \sigma^{-1}\phi\left(\frac{\log x_j - (\theta_j + \log S)}{\sigma}\right), \end{aligned} \tag{5}$$

where L_j is the likelihood contribution from the jth locus, Φ is the cumulative distribution function for the standard normal distribution and ϕ is the probability density function of the standard normal distribution. The first product sign, $\prod_{j\in J^C}$, collects the likelihood contribution of the loci that dropped out because $\Phi(\log t - (\theta_j + \log S)/\sigma)$ is the probability of observing a value less than $\log t$ in a $N(\theta_j + \log S, \sigma^2)$ distribution. The second product sign, $\prod_{j\in J}$, collects the likelihood contribution from the loci that did not drop out because $\sigma^{-1}\phi(\log x_j - (\theta_j + \log S)/\sigma)$ is the probability of observing the value $\log x_j$ in a $N(\theta_j + \log S, \sigma^2)$ distribution.

For a sample $\{x_j\}_{j\in J}$, the likelihood in Eq. (4) can be optimised numerically using the `optim` functionality in `R` [17] to obtain the estimate $\log \hat{S}$. Note, that if we have a full profile, ($J^C = \emptyset$), then the optimum of Eq. (4) is $\log \hat{S} = r^{-1}\sum_{j=1}^{r}(\log x_j - \theta_j) = r^{-1}\sum_{j=1}^{r}\log x_j$. In other words, for a full profile, the log of the signal strength is the average of the log peak heights because the sum of the locus balances is 0. Also, note that at least two loci are required because both $\log S$ and σ^2 must be estimated.

If the information about truncation is ignored, then the crude estimator

$$\log \hat{S}_{crude} = \frac{1}{r-k}\sum_{j\in J}(\log x_j - \theta_j) \tag{6}$$

can be used, where $k = |J^C|$ is the number of loci dropped out. The crude estimator is expected to be greater than the likelihood estimator because it does not incorporate knowledge of the loci dropped-out and the estimate is decreased because the peaks dropped-out are known to be smaller than 50 RFU. In Fig. 6, the signal strength estimator based on optimising the likelihood in Eq. (4) is compared to the crude estimator in (5) using all the data from profiles with at least two loci not dropped out. This figure shows that the crude estimator in Eq. (5) is greater than the likelihood based estimator in Eq. (4).

Estimators of truncated normal distributions are treated in [18], but locus imbalances make things complicated, which is why we use the numerical optimisation.

Optimising Eq. (4) makes it possible to estimate the signal strengths, $\hat{S}$, for all samples with at least two loci not dropped out. Only these samples with at least two loci not dropped out are used. In principle, the crude estimator Eq. (5) could be used, but as described previously and shown in Fig. 6, this would result in too large signal strengths for samples with only one locus. Another option would be to estimate the overall variance σ^2 such that only one observation would be needed to estimate the one parameter S. As shown in Fig. 7, the variance for low signal strengths is probably too large to obtain a reasonable overall estimate.

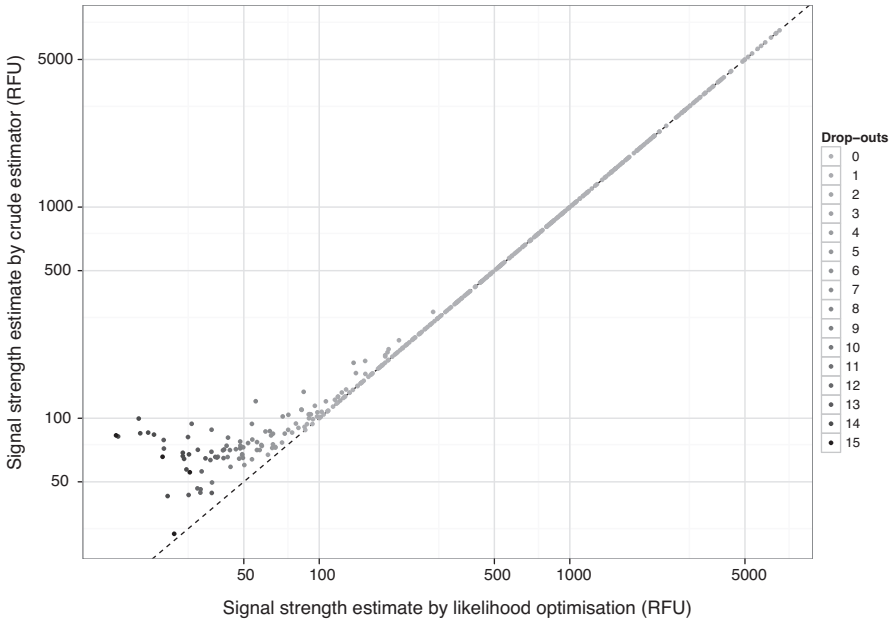

Fig. 6. Comparison of the signal strength estimator based on the likelihood Eq. (4) and the crude estimator Eq. (5) based on all data with at least two loci not dropped out. The line has slope 1 and intercept 0 corresponding to a 1:1 correlation. The crude estimator is expected to be greater than the likelihood estimator (see the text for the arguments), which is supported by this figure (because the points are above the line).

In Fig. 8, the correlation between the DNA concentration and the signal strength given the number of PCR cycles is shown. In Fig. 9, the correlation between signal strength and the proportion of loci dropped out is depicted.

2.5. Modelling drop-out probability

As done in other studies, e.g. [15,16], logistic regression [19,20] of the probability of drop-out was performed. Possible explanatory variables considered were `Experiment`, `LogSignalStrength` ($\log S_i$), `Cycles` (28, 29, or 30 PCR cycles), `Locus`, `Dye` and `FragmentSize`.

We performed backwards model selection using the Bayesian Information Criterion [21] (BIC) to select the best model. The initial model consisted of all first order effects and second order interactions (for example to allow the effect of signal strength to depend on the number of PCR cycles).

3. Results

3.1. Model for drop-out probability

As described in Section 2.5, a logistic regression was used to estimate the probability of drop-out. The resulting model was that

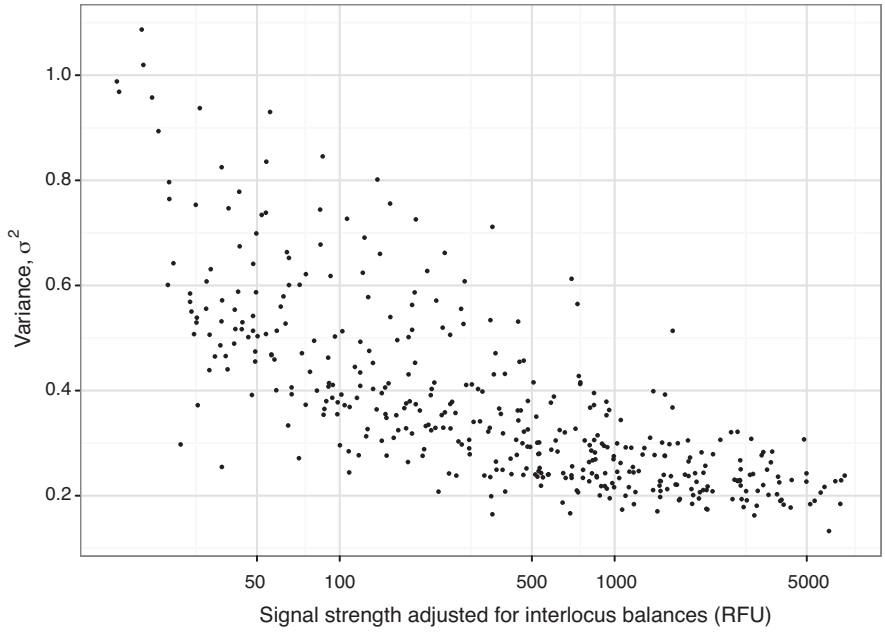

Fig. 7. The variance, σ_i^2, for each sample given the signal strength, S_i, based on optimising Eq. (4). The variance, σ_i^2, decreased with the signal strength.

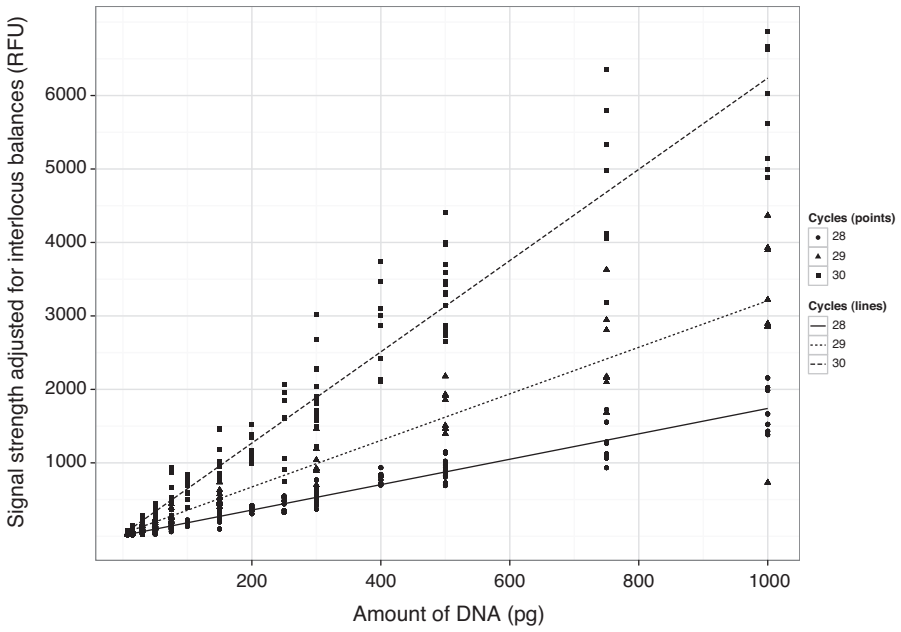

Fig. 8. The correlation between the DNA amount and the estimated signal strength (as explained in Section 2.4) given the number of PCR cycles. The lines are linear regression lines for each of the PCR cycles.

the drop-out probability is best described by an effect of `LogSignalStrength` ($\log S_i$), `Cycles`, `FragmentSize` and an interaction effect between `LogSignalStrength` and `Cycles` such that the effect of signal strength varies with the number of PCR cycles.

The drop-out probability given signal strength for fragment size 150 bp is shown in Fig. 10.

The corresponding signal strength given a drop-out probability for fragment sizes 150 and 300 bp is shown in Fig. 11. Table 1 shows the figures.

3.2. Model validation

To validate the model, an Hosmer–Lemeshow's test [20] and a bootstrap validation [22] of the receiver operating characteristic (ROC) were performed.

In total, the dataset contained 6565 rows (one row per peak). Because of this relatively high number of observations, 50 groups were chosen for the Hosmer–Lemeshow's test. The resulting test statistic was $X^2 = 38.7$, resulting in a non-significant result ($p = 0.83$), meaning that it could not be rejected that the data could be explained by the model.

For a dataset with n samples, the bootstrap procedure was as follows: n samples were randomly chosen *with* replacement and used to fit the model. The samples from the dataset that were not chosen were then used to validate the model. This was repeated 1000 times calculating the receiver operating characteristic (ROC). More specifically, the area under the ROC curve (AUC), the sensitivity, and specificity were used as validation statistics. The value of sensitivity and specificity were taken at the cutoff, which was the point, where both were highest with equal weight

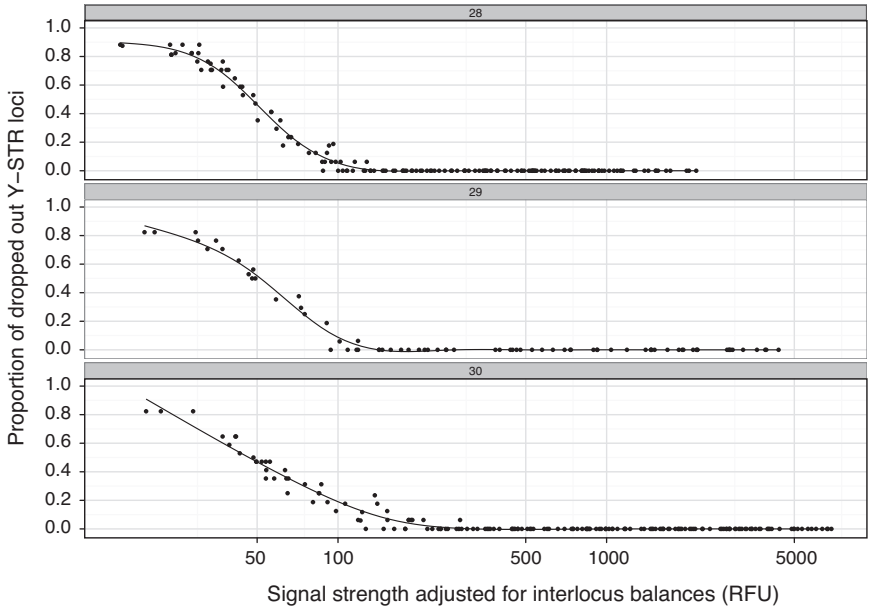

Fig. 9. The proportion of dropped out Y-STR loci given signal strength $\theta_j + \log S_i$.

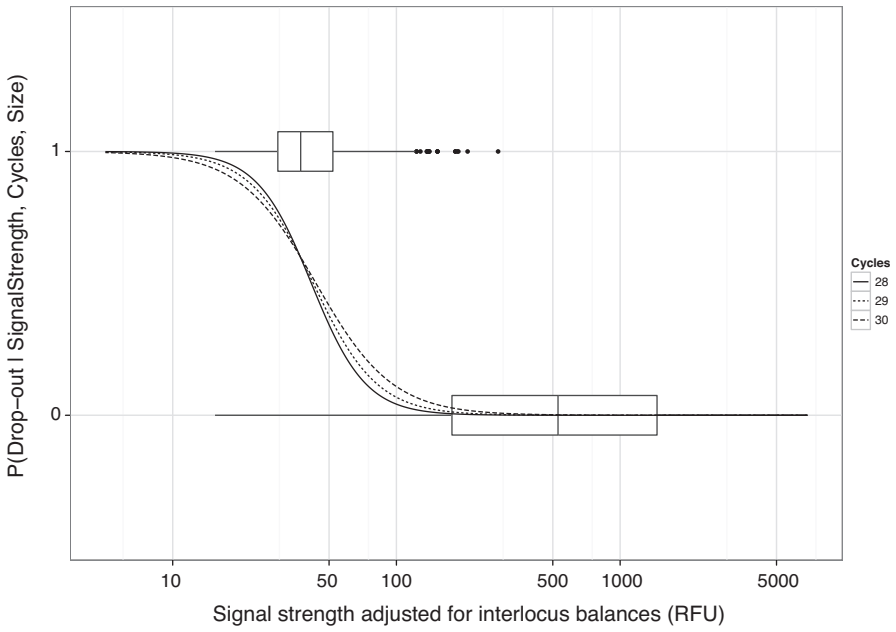

Fig. 10. Drop-out probabilities given signal strengths for a fixed fragment size of 150 bp.

(meaning that both are treated as equally important, which may not always be the case).

Fig. 12 shows the results of the receiver operating characteristic (ROC) analyses of the 1000 bootstrap realisations. As seen, the results of the ROC analyses did not contradict the proposed model being sufficient to describe the data.

4. Discussion

The result of our investigations indicated that the drop-out probability can be sufficiently described by $\log \hat{S}$ (where $\hat{S}$ is an estimate of the signal strength in a profile), the number of PCR cycles, and fragment size. Note, that the locus balances are incorporated in the calculation of $\log \hat{S}$.

The effects of experiments were not sufficiently strong to be included as a covariate at the model selection, meaning that no significant day-to-day effect was observed. It would be interesting to investigate whether differences in kit-lot number have effect on the parameters under study. Unfortunately, the lot numbers were not recorded.

Going back to the motivating example in Section 1.1, our analysis showed, based on Table 1, that for 30 PCR cycles with a 150 bp allele, the probability of drop-out was 1:5000 corresponding to a rough estimate of the probability of null alleles at a signal strength of $S = 1249$ RFU. This means that the probability of a null allele is higher than that of drop-out at 4000 RFU and that the probability of drop-out is higher than that of a null allele at 75 RFU.

We have developed a model suitable for pristine DNA without degradation. The model can be extended to encompass degraded Y chromosomal DNA similar to the way [23] models degraded autosomale DNA.

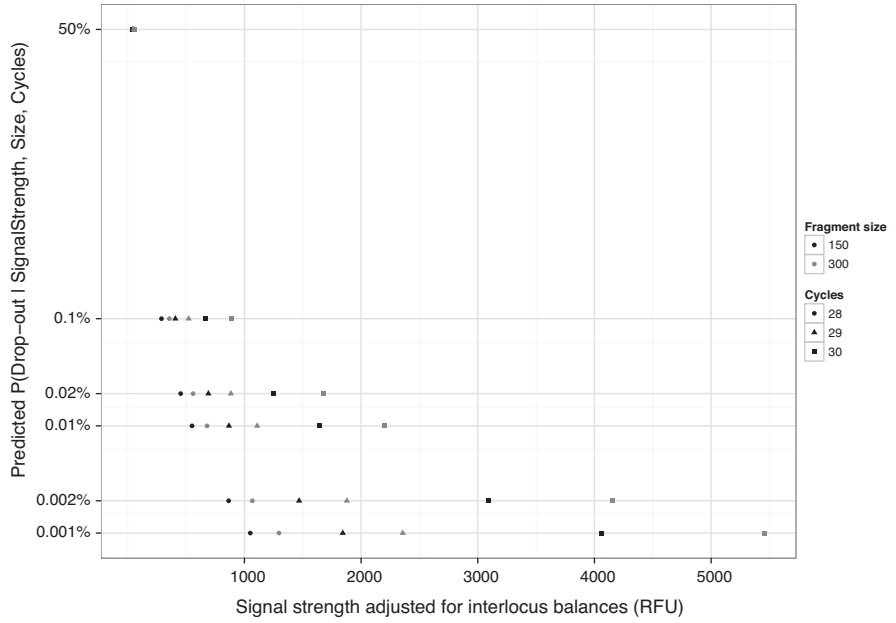

Fig. 11. Plot of signal strength given a drop-out probability for fixed fragment sizes 150 and 300 bp. See Table 1 for a table of values used to construct this plot.

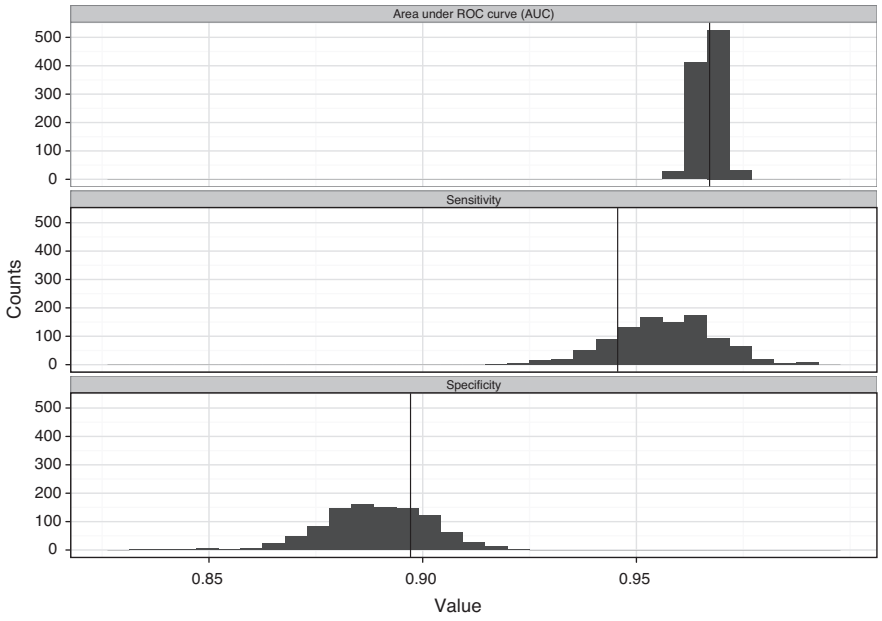

Fig. 12. Realisations of the area under curve (AUC), sensitivity and specificity from the ROC analyses of 1000 bootstrap samples. The vertical lines are the values obtained when both fitting and validating the model using the entire dataset.

4.1. Locus balances

As already shown in Fig. 4, the Yfiler kit is not well balanced. The imbalance seems to be independent of the DNA concentration (not shown). This makes it difficult to make a good model for estimating signal strength.

In Section 2.3, we described a model to estimate the locus balances shown in Fig. 5. We will now describe a more advanced model for estimating the signal strength. The idea is that loci with smaller variance contribute with more information to the estimation of the signal strength.

Going back to Fig. 4, not all loci have the same variance meaning that they each contribute with a different amount of information. Let ϕ_j^2 be the variance of the jth locus' proportion of the sum of peaks heights (resembles the width of the boxes in Fig. 4). As in Eq. (1), the full profiles are used to estimate the θ_j's and ϕ_j^2's by using the model

$$\log x_{ij} \sim N(\theta_j + \log S_i, \phi_j^2). \tag{7}$$

The estimated ϕ_j^2's are depicted in Fig. 13. The estimated θ_j's and ϕ_j^2's are then assumed known when used in the model for estimating signal strength, such that

$$\log x_{ij} \sim N_{\log t}(\theta_j + \log S_i, \phi_j^2 \sigma_i^2), \tag{8}$$

where x_{ij} is the peak height at the jth locus for the ith sample, θ_j is the locus balance for the jth locus and S_i is the signal strength for

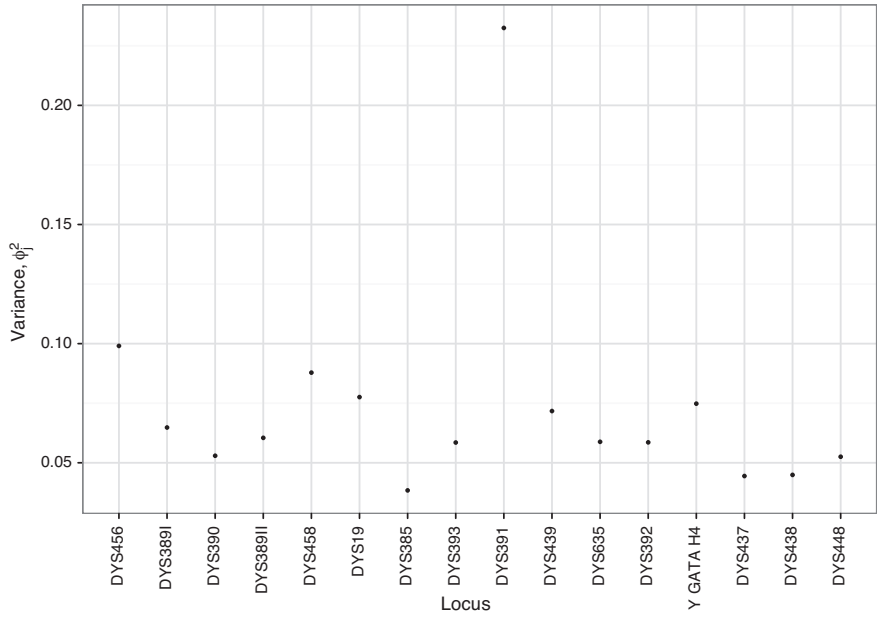

Fig. 13. Estimation of the variance, ϕ_j^2, using the model Eq. (6). The values can be compared to the width of the boxes in Fig. 4. Loci with a large box width in Fig. 4 also have a large variance, ϕ_j. One example of this is the DYS391 locus.

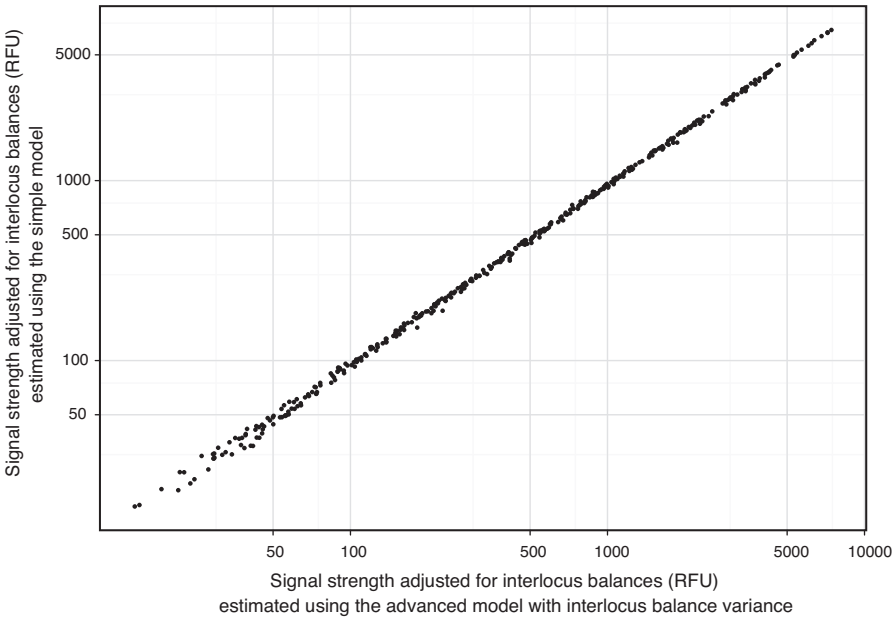

Fig. 14. Comparison of the signal strength estimation using the advanced model Eq. (7) and the simple model Eq. (2). Each point represents the estimated signal strength of a sample using both the advanced and simple model.

the *i*th sample. As seen, Eq. (7) is an extension of Eq. (2). The likelihood, which for Eq. (2) was Eq. (4), to be optimised is then

$$L(\log S, \sigma^2; \{x_j\}_{j\in J}) = \prod_{j\in J^C} \Phi\left(\frac{\log t - (\theta_j + \log S)}{\phi_j \sigma}\right) \times \prod_{j\in J} (\phi_j\sigma)^{-1} \phi\left(\frac{\log x_j - (\theta_j + \log S)}{\phi_j \sigma}\right).$$

The results for the two different ways of estimating signal strength are shown in Fig. 14. As seen, the results obtained using the advanced model are quite similar to the results obtained using the simpler model. This does not mean that the variance of the interlocus balances is not important, merely that it is probably difficult to model.

References

[1] P. Gill, C. Brenner, et al., DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs, Forensic Sci. Int. 124 (2001) 5–10.
[2] L. Gusmao, J. Butler, et al., DNA Commission of the International Society of Forensic Genetics. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, Forensic Sci. Int. 157 (2006) 187–197.
[3] L. Roewer, Y chromosome STR typing in crime casework, Forensic Sci. Med. Pathol. 5 (2009) 77–84.
[4] M. Prinz, K. Boll, H. Baum, B. Shaler, Multiplexing of Y chromosome specific STRs and performance for mixed samples, Forensic Sci. Int. 85 (1997) 209–218.
[5] N. Morling, R. Allen, et al., Paternity Testing Commission of the International Society of Forensic Genetics: recommendations on genetic investigations in paternity cases, Forensic Sci. Int. 129 (2002) 148–157.
[6] P. Gill, C. Brenner, J. Buckleton, A. Carracedo, M. Krawczak, W. Mayr, N. Morling, M. Prinz, P. Schneider, B. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101.
[7] L. Roewer, M. Kayser, P. de Knijff, et al., A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males, Forensic Sci. Int. 114 (2000) 31–43.
[8] M. Krawczak, Forensic evaluation of Y-STR haplotype matches: a comment, Forensic Sci. Int. 118 (2001) 114–115.
[9] C.H. Brenner, Fundamental problem of forensic mathematics—the evidential value of a rare haplotype, Forensic Sci. Int. Genet. 4 (2010) 281–291.
[10] J. Buckleton, M. Krawczak, B. Weir, The interpretation of lineage markers in forensic DNA testing, Forensic Sci. Int. Genet. 5 (2011) 78–83, Haploid DNA markers in Forensic Genetics.
[11] J.M. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, second ed., Academic Press, 2005.
[12] B. Budowle, X. Aranda, et al., Null allele sequence structure at the DYS448 locus and implications for profile interpretation, Int. J. Legal Med. 122 (2008) 421–427.
[13] L. Roewer, M. Krawczak, S. Willuweit, et al., Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes, Forensic Sci. Int. 2–3 (2001) 106–113.
[14] S. Willuweit, L. Roewer, Y chromosome haplotype reference database (YHRD): update, Forensic Sci. Int. Genet. 1 (2009) 83–87.
[15] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Estimating the probability of allelic drop-out of STR alleles in forensic genetics, Forensic Sci. Int. Genet. 3 (2009) 222–226.
[16] T. Tvedebrink, P.S. Eriksen, M. Asplund, H.S. Mogensen, N. Morling, Allelic drop-out probabilities estimated by logistic regression—further considerations and practical implementation, Forensic Sci. Int. Genet. 6 (2011) 263–267.
[17] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0.
[18] T. Persson, H. Rootzen, Simple and highly efficient estimators for a Type I censored normal sample, Biometrika 64 (1977) 123–128.
[19] A. Agresti, Categorical Data Analysis, second ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.
[20] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, John Wiley & Sons, Inc., Hoboken, New Jersey, 2000.
[21] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[22] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[23] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out, Forensic Sci. Int. Genet. 6 (2011) 97–101.
[24] M. M. Andersen, A. Caliebe, A. Jochens, S. Willuweit, M. Krawczak. Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. Forensic Science International: Genetics, in Press, http://dx.doi.org/10.1016/j.fsigen.2012.11.004.